DOCUMENT RESUME

ED 322 725 FL 018 726

AUTHOR Morrow, Keith
TITLE Evaluating Communicative Tests.
PUB DATE Apr 90
NOTE 9p.; Paper presented at the Annual Meeting of the Regional Language Centre Seminar (Singapore, April 9-12, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Communicative Competence (Languages); *English (Second Language); *Evaluation Criteria; Foreign Countries; *Language Tests; Second Language Instruction; *Test Construction; Testing
IDENTIFIERS *Certificates in Communicative Skills in English; *Great Britain

ABSTRACT

The influence of the communicative approach to second language teaching on the design of the new Certificates in Communicative Skills in English tests is examined, and the evaluation of the tests in practice is discussed from that perspective. It is proposed that a number of characteristics of the new series should relate directly and consciously to an underlying concept of what a good test should be. They: (1) test single skills; (2) are direct measures of performance; (3) are task-based and criterion-referenced; and (4) reflect and encourage good classroom practice. Issues to be addressed in evaluating design features are outlined, including: (1) why specific abilities, not overall proficiency, are tested; (2) whether or not there is a conflict between authenticity and reliability; and (3) test validity. Alternatives to traditional methods of language testing are needed. Sources for alternative ideas are found in chaos theory, and in work on the validation of experiential research. These ideas move the center of attention in language testing away from the conventional focus on the acquisition and analysis of hard data to include a concern with those aspects of the nature of language itself that may not be susceptible to such procedures and with the effects that tests and testing have on the consumers. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Presented at the 1990 Regional Language Centre Conference

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
K. Morrow
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED322725

EVALUATING COMMUNICATIVE TESTS

by Keith Morrow

1. The Context:CCSE

In 1976, while working at the Centre for Applied Language Studies at the University of Reading, I was commissioned by the Royal Society of Arts to undertake a feasibility study into the development of a series of English language examinations based on the ideas about "communicative" language teaching which were then taking shape at Reading and elsewhere. 1. The outcome of this study was a series of examinations called the Communicative Use of English as a Foreign Language run by the RSA between 1980 and 1988, and subsequently run jointly by the RSA and the University of Cambridge Local Examinations Syndicate. I have recently completed a review of these examinations for Cambridge/RSA, and from November 1990 they are to be re-launched as Certificates in Communicative Skills in English (CCSE).

It will be clear that the origins of these examinations are to be found in a particular, though now widespread, view of what language, language teaching and language testing is about. In this paper I want to consider how this view has affected the design of the examinations, and then look at the question of the evaluation of the examinations in practice from the same perspective.

A number of characteristics of the new series of examinations relate directly and consciously to an underlying construct of what a "good" test ought to be.

1.1 Single skills

The examinations in fact consist of a suite of free-standing modules in Writing, Reading, Listening and Oral interaction. In each skill area, new tests are set at 4 levels for each series of the examination, and candidates are able to choose which modules at which level they wish to enter at any time. This structure reflects the experience of language teachers that the performance of students is not uniform across skill areas.

FL 018 726

Notes

1 Techniques of Evaluation for a Notional Syllabus RSA (London) 1977. Now available from EFL Dept, UCLES, 1 Hills Rd, Cambridge, CB1 2EU U.K.

BEST COPY AVAILABLE

1.2 Tests of Performance

The tests are designed to be direct measures of performance. The justification from this again derives largely from an educational perspective on how tests should affect what is done in the classroom. In a "communicative" classroom, the focus of activities is (in simple terms) "doing something" through the language. We wanted to encourage this in washback terms, by designing tests which shared the same focus.

1.3 Task Based

More specifically, communication through language involves the participants in carrying out "tasks" in the production or comprehension of language in order to arrive at a shared understanding. This "negotiation of meaning" is most obviously a feature of face-to-face interaction, but it underlies all purposeful use of language whether in reading, writing, listening or speaking. The most striking implication of adopting this perspective on the design of a language test is the overriding importance of authenticity both of text (as input) and of task (in processing this input).

1.4 Criterion-referenced

The essential question which a communicative test must answer is whether or not (or how well) a candidate can use language to communicate meanings. But "communicating meanings" is a very elusive criterion indeed on which to base judgements. It varies both in terms of "communicating" (which is rarely a black and white, either / or matter) and in terms of "meanings" (which are very large and probably infinite in number). In other words, a communicative test which wishes to be criterion-referenced must define and delimit the criteria. This is a major undertaking, which for CCSE has led to statements for each of the 4 levels in each of the 4 skill areas of the tasks and text types which the candidates are expected to handle as well as (crucially) the degree of skill with which they will be expected to operate.

1.5 To reflect and encourage good classroom practice

Reference has already been made above to the educational effect of testing through the promotion of positive washback into the classroom. In the case of CCSE, this is a major concern underlying the design of the tests; indeed in many ways the tests themselves have drawn on "good" classroom practice in an attempt to disseminate this to other classrooms. This conscious feedback loop between teaching and testing, in terms not only of content but also of approach, is a vital mechanism for educational development.

2. Evaluating Tests of this Kind

It will be clear from the preceding section that there is a conscious and deliberate rationale underlying the construction of the CCSE tests. However, a rationale can be wrong and some of the bases on which the tests are constructed may appear to be assertions of what is believed, rather than outcomes of a process to determine what is right. Most significantly a "professional" language tester might perhaps wish to investigate the justification for the following design features.

2.1 Why tests of specific abilities not overall proficiency?

The idea of a single measure of overall competence or proficiency in a language is an attractive one for test designers and educational administrators. For administrators, the convenience of a single score or rating is obvious, and tests and exams as diverse as TOEFL, the Cambridge First Certificate, and the new Cambridge/ British Council IELTS all manage to provide this. Why shouldn't CCSE? For test designers and researchers an "overall proficiency" model may also be attractive, not least for the scope it offers for writing papers reporting sophisticated investigations into the existence of this underlying "g" and the factors which may or may not contribute to it.

The CCSE scheme minimises the problem of weighting different factors by reporting performance in terms of each skill area individually. However, since within each skill area decisions have to be made about the contribution of individual tasks involving specific text types to the overall performance. there will still be scope for investigations of what makes up the underlying "l","r", "s" and "w" factors in the listening, reading, speaking and writing tests.

The main justification for the apparent complexity of the structure of CCSE is not, in fact, to be found in the literature on language testing. Rather it is to be found in a deliberate attempt to provide on educational grounds an examination system which allows candidates to gain certification for what they can do without being penalised for what they cannot do. This is a stance which reflects a philosophical rather than an empirical starting point. Nonetheless, it is a common experience of teachers that many students have differential abilities in different language skill areas; it is important both practically and educationally that this is recognised.

2.2 Is there a conflict between authenticity and reliability?

The short answer to this question is probably "yes". In terms of test design, CCSE clearly needs to address itself to (at least) two areas of potential problem caused by the focus on the use of authentic tasks and texts. The first is the question of consistency of level of the tasks / texts used both within a particular test, and across tests in the same skill area in different series; the second is the question of consistency of judgement in the evaluation of performance, particularly in the oral interaction and writing tests.

In the implementation of the CCSE scheme, rigorous safeguards are in place to take account of these potential problems. A thorough moderating procedure is undertaken to scrutinise and trial tasks and papers to ensure consistency of content; and assessors for both the writing and oral tests are trained, briefed and monitored.

Yet still, it has to be said that the conflict remains. There are steps that could be undertaken to improve the reliability of the CCSE tests; but they would conflict directly with the authenticity criterion. Once again, it seems that in test design terms, the way that is chosen reflects a basic educational philosophy. From one standpoint, reliability is crucial; authenticity can be brought in to the extent that it is possible, but it remains secondary. From another, the essential characteristic is authenticity; while recognising that total authenticity of task can never be achieved in a testing (or teaching) situation, every effort is made to focus on it. Reliability is not ignored and indeed every effort is made to ensure it receives due attention, but in the final analysis it is not the overriding factor in the design of the test.

It seems unlikely that in principle this conflict can be resolved. What is perhaps more important is to investigate how in practice the implementation of communicative exams like the CCSE proceeds.

2.3 How valid are the tests?

The third area of concern clearly focuses on validity. In the literature, language proficiency tests are investigated in terms of construct, content and concurrent validity, and face validity is also considered - sometimes as an afterthought, sometimes if nothing else can be found, but because it cannot be "measured", sometimes rather disparagingly. Some specific purpose tests are also evaluated in terms of predictive validity.

It will be clear from the preceding discussion that there is a "construct" to the CCSE scheme; the relationship between the construct and the content of any particular set of tests is open to empirical investigation - but unfortunately the construct itself is probably not. Similarly, in terms of content, the relationship between the specifications and the content of any particular set of papers can be investigated; but the more fundamental question of

how far the specifications reflect the "real world" is not a matter of straightforward analysis. Concurrent validity is largely irrelevant because there are few other tests available against which CCSE can be measured. Face validity can be claimed - and claimed to be extremely important - but not proved. It may seem that CCSE should be open to investigation in terms of predictive validity since it is in essence making claims about the ability of candidates to carry out "real world" tasks. If they pass the test and can in fact carry out these tasks in the real world then the test may be said to have "predicted" this.

However, a moment's thought will show that this is in fact an impossible requirement. How would a researcher judge whether or not tasks had been "carried out" in the real world? The only way would be by evaluating performance on individual instances of the task - in which case all the problems of specification, reliability and generalisability which arise for CCSE would arise again. In a very real sense, the test would be validating itself against itself.

3. Epistemological Traditions

In the preceding section we have seen how in three very basic respects, the design of a communicative test is based on factors which go beyond, or are not susceptible to, conventional language testing research: the overall design is founded on educational rather than testing requirements; reliability is secondary on construct grounds to authenticity; and the fundamental validity is not open to straightforward investigation.

This situation seems to raise some rather interesting questions about the kind of research which is appropriate to the investigation of language tests.

3.1 Language testing as pure science

Since the 1960's and the development of the "scientific" approach to testing, the predominant model of research has been one based on the precise measurement of data collected through empirical means. In other words, language testing has been seen as a branch of pure science, based essentially upon the twin concepts of quantification and analysis. In many ways this model has been extremely productive - at least of articles, books and PhD theses. But in terms of actually being able to make reliable, valid and comprehensible statements about what it is that tests measure, how this relates to language use, and how performance on tests relates to the real world, one might be tempted to echo the comments of a Nobel Prize winning economist about his own subject:

"In no field of empirical enquiry has so massive and sophisticated a statistical machinery been used with such indifferent results." 2

It seems to me that the reason for this sad state of affairs may well lie with the very notion of scientific rigour as it is generally understood, and the orthodoxy that this has imposed on our field.

"Under normal conditions the research scientist is not an innovator but a solver of puzzles, and the puzzles upon which he (sic) concentrates are just those which he believes can be both stated and solved within the existing scientific tradition" 3

At its best, this existing scientific tradition has encouraged researchers into what has become known as "McNamara's fallacy" of "making the measurable important instead of making the important measurable (or at least discernible)".4 At its worst, it might tempt us to paraphrase Oscar Wilde's epigram about the English upper classes and their fondness for fox-hunting. Wilde spoke of "The English country gentleman galloping after a fox - the unspeakable in full pursuit of the uneatable." Language testing researchers may not be unspeakable; but they may well be in pursuit of the unmeasurable.

Elsewhere,5 I have suggested that it may be time to propose a distinction between language testing researchers (who follow the existing orthodoxy) and researchers into language testing who would be prepared to adopt equally rigorous but rather different ways of looking at the area.

Support for alternative ways of conducting research into language testing seems to me to be available from two very different sources.

The first is the recent development (or at least the recent dissemination) of the ideas behind chaos theory. An exposition of this theory would be out of place here 6 (and in detail beyond my present understanding of it). But I find a set of ideas leads to

---

2 W. Leontiev quoted in Chaos by J. Gleick Cardinal 1988

3 Kuhn T.S. The Essential Tension: Selected Studies in Scientific Tradition and Change. University of Chicago 1977 Quoted in Gleick 1988 (See note 2 above)

4 Quoted in Assessing Students: How Shall We Know Them by D. Rowntree. Open University 1977

5 See Evaluating Tests of Communicative Performance in Innovation in Language Testing ed. M. Portal Nelson/NFER 1987

6 For an excellent popular account, see Gleick 1988 (See note 2 above)

the insight that conventional science finds it impossible to make a definitive statement about the length of a coastline (because of the problems of scale: the larger the scale, the longer the length because the more "detail" is included), or a firm prediction about the temperature of a cup of "hot" coffee in a minute's time (because of the variability of convection), let alone what the weather is going to be like next week (because of the cumulative effect of a whole range of unpredictable and in themselves "trivial" events) an extremely powerful heuristic in thinking about language and language testing. Perhaps the key concept is "sensitive dependence upon initial conditions" - a way of saying that in looking at the world, everything depends on precisely where you start from. Nothing could be more appropriate for our field.

The second source of alternative ideas for the investigation of language testing comes from work carried out in the validation of experiential research. An extremely lucid account of this is given in Heron 1982. 7. Experiential research is concerned above all to establish the quality or nature of the learning experience which participants have undergone. In this, it seems to me to relate very closely to important areas of concern for research into language learning and language testing.

Heron sets out a number of categories of validity which should be met by research. Empirical and conceptual validity can be related easily enough to categories which are familiar in our field; but his third category ethical validity perhaps opens up new areas. How far does existing research into language testing concern itself with the ethics of the test?

"Ethical validity has two aspects. Firstly is the research relevant to basic human concerns? Is it committed to values that make a difference to the quality of life for people now and in the future?...Secondly, do we behave morally while doing and applying the research...And do we deploy the results in ways that respect the rights and liberties of persons?" (Heron: 1982 p.1)

There are many questions raised here about the design and implementation of language tests, but perhaps the most obvious area of ethical involvement is the question of washback. It seems to me that a test which imposes (overtly or covertly) an impoverished or unrealistic classroom regime on students preparing for it is indeed "making a difference to the quality of life for people now and in the future" (though a difference of the wrong sort). This reinforces my view 8 that an important area of investigation in considering the validity of a test is an investigation of the classroom practice which it gives rise to.

---

7 John Heron Empirical Validity in Experiential Research Dept Adult Education, University of Surrey, Guildford Surrey GU2 5XH

8 See reference in Note 5 above.

A large part of Heron's paper is taken up with considering "Procedures for Distinguishing between the Veridical and the Illusory" in the context of experiential research. Again, this seems a rich field to harvest in considering the validity of language tests. This is not the place to consider all of these in detail, but one in particular is worthy of note. It is the principle of "authentic collaboration" between all the participants in a research project, breaking down "the traditional distinction between the role of the researcher and the role of the subject". As Underhill 9 points out, the use of the label "subjects" is a widespread but de-humanising feature of most current work in language testing. In principle, the job of finding out what somebody is able to do in a language, or what the effectiveness of a particular test or test procedure is, might be greatly facilitated and would be much more "ethical" if the "subjects" were themselves genuinely involved in the process. Working out how to do this is of course another matter. But it does seem to point a new and very interesting direction for language testing and its associated research.

Underlying these two sources of new input to research on language testing, then, are ideas which move the centre of attention away from the conventional focus on the acquisition and analysis of "hard" data, to include a concern with those aspects of the nature of language itself which may not be susceptible to such procedures, and with the effects which tests and testing may have on the consumers. Testing and research which reflects this move will no longer be concerned simply to measure; rather it will establish a framework which will permit informed and consistent judgements to be made.

It would be unrealistic to claim that the CCSE scheme meets all the criteria (implicit and explicit) set out in this last section. But it is perhaps reasonable to claim that it illustrates the legitimacy of asking questions about tests other than those which language testing researchers conventionally ask. The next step is to find some answers.

Keith Morrow
Education Advisor
Bell Educational Trust
Cambridge GB

9 See N. Underhill Testing Spoken Language Cambridge University Press 1988